applied sciences

MDPI

Article

Determination of the Features of the Author's Style of A.S. Pushkin's Poems by Machine Learning Methods

Vladimir Barakhnin [1,2], Olga Kozhemyakina [1,*] and Irina Grigorieva [1]

[1] Federal Research Center for Information and Computational Technologies, 630090 Novosibirsk, Russia; bar@ict.nsc.ru (V.B.); igriva@list.ru (I.G.)
[2] Department of Information Technologies, Novosibirsk State University, 630090 Novosibirsk, Russia
[*] Correspondence: olgakozhemyakina@mail.ru; Tel.: +7-913-946-22-80

Abstract: This paper presents the study of the author's style of A.S. Pushkin based on the comparison of his poetic texts with the texts of contemporary poets. The purpose of this study is to determine the features of the author's style of A.S. Pushkin using machine learning methods. This paper describes the construction of several classifications based on different groups of features, as well as the classification based on a combined set of features from different groups. The quality of all constructed classifications is also analyzed; special attention is paid to the interpretation of the neural network solution and the identification of features of the author's style.

Keywords: stylometry; poetry; text model; features of the author's style; machine learning; neural network

check for updates

Citation: Barakhnin, V.; Kozhemyakina, O.; Grigorieva, I. Determination of the Features of the Author's Style of A.S. Pushkin's Poems by Machine Learning Methods. *Appl. Sci.* **2022**, *12*, 1674. https://doi.org/10.3390/app12031674

Academic Editor: Valentino Santucci

Received: 22 December 2021
Accepted: 31 January 2022
Published: 6 February 2022

1. Introduction

The stylometry is currently the interdisciplinary field combining literary stylistics, statistics, machine learning and computer linguistics to study the style of documents with various purposes. Currently, the stylometry is used to determine the authorship, to describe the features of the author's style, to detect plagiarism or falsification of various documents, articles, traditional letters, emails and short messages on social networks, and, finally, literary works and their fragments, as well as scientific papers. The stylometry is constantly being improved, although the universal methods leading to a reliable determination of authorship have not yet been developed for most tasks. Nevertheless, in recent years, the methods of stylometry have become actively used in the protection of information, as described in [1–3], and criminology [4].

The modern stylometry is more connected with the analysis of printed texts presented in electronic form, which is due to the active digitalization of society. The need to solve a significant number of problems of stylometry has led to the expansion of the usage of computer methods. Currently, the effectiveness of computational linguistics methods and machine learning methods in the problems of stylometry is already difficult to dispute, these methods are being developed by groups of researchers around the world.

The stylometry in the context of attribution of authorship assumes the identification of features of the author's style that can be quantified. This work presents the study of the author's style based on poetic texts in Russian. The task is to determine the features of the author's style of A.S. Pushkin, as well as poets of the Pushkin era, using machine learning methods. These methods require the data to compare and to identify the features that characterize the author's style, so the problem was formulated as a binary classification problem, in which two classes were distinguished: the poems by A.S. Pushkin, and poems by other poets of the Pushkin era: K.N. Batyushkov, E.A. Boratynsky, P.A. Vyazemsky, N.I. Gnedich, D.V. Davydov, A.A. Delvig, and V.A. Zhukovsky. It was decided to use only poems as data (the work belonged to this genre on the basis of the formal characteristics given by the compilers of the collected works).

Here, are the main terms used in this article in relation to Russian versification.

A poem is a work written in verse, mostly of a small volume (unlike prose) and mostly lyrical [5].

A verse is a literary speech, phonetically divided into the separate segments (each of them is also called a verse), which are perceived as comparable and commensurate [6]. In comparison with prose, the division of verse has two features:

1. In prose, the division of text is determined only by syntactic pauses; in verse, the articulating pauses may not coincide with syntactic ones;
2. In prose, the allocation of articulating pauses is largely arbitrary; in verses, it is firmly set.

As noted in [5], the division into verses is usually marked by the graphic design of the text and is often accompanied by rhyme or other phonetic features. A means to emphasize the comparability and commensurability of verses is the meter—the alternation of strong and weak points within the verse—but it may also be absent, for example in purely tonic or free verse.

The poems as data for classification have a certain specificity: as these are small works, the usage of methods that take into account the frequency of various elements of the verse or compression methods may not give sufficiently accurate results. Therefore, it was decided to abandon the usage of compression methods immediately, but the features describing the frequency of elements form the basis of this work. The metrorhythmic characteristics of a verse can provide additional useful information for classification.

2. Definition of Feature Groups

The attribution tasks and the definition of the author's style are similar in many ways; they require the determination of the style elements inherent in a particular author, but the attribution tasks are often formulated somewhat differently, for example as described in [7]. Within the framework of this work, the classification serves as the basis for the analysis of features in order to highlight the characteristics of the author's style.

From the point of view of philology, the individual (written) style is a complex concept reflecting the socio-historical nature, ethnic, psychological, moral, and ethical characteristics of the author [8]. The various authors identify the different levels of text analysis [8,9] that can be used in stylometry and, upon combining them, we find the following set of levels:

- The phonetic level takes into account the peculiarity of intonation and melodica. The melodica describes the number of syllables in words, the repetitions of vowels and consonants to enhance expressive power and euphony. In poetic works, the melodica is aimed at ensuring the rhythm and the harmony.
- The punctuation level reveals the peculiarities of the author's usage of punctuation marks and characteristic errors.
- The spelling level reveals the characteristic errors in the spelling of words.
- The lexical level describes the author's vocabulary, the features of the usage of words, phrases and stable expressions, as well as typical repetitive parts of sentences; a tendency to use rare and foreign words, synonyms, antonyms, paronyms, neologisms, as well as words denoting certain concepts.
- The syntactic level describes the features of sentence construction, the prevailing types of sentences (affirmative, interrogative, exclamation), the tendency to use complex or incomplete, as well as elliptical sentences, types of syntactic coherence, word order, and sentence length.
- The stylistic level is responsible for the genre, plot, and general structure of the text, as well as characteristic artistic means.

As only published texts are considered in this work, the presence of spelling and punctuation errors is excluded. In the work [9], it is noted that the author's style usually refers to the features of the text at the highest levels, namely syntactic, lexical, and stylistic levels. It should be noted that these levels are widely used in expert analysis, but for formal computational analysis, these groups are quite complex, so most researchers use phonetic, as well as punctuation, lexical, and syntactic levels for analysis.

Within the framework of this work, the groups of features are identified that actively influence the definition of authorship and characterize the author's style the most. The following groups of features were identified:

- the distribution by parts of speech and relationships;
- the punctuation mark frequencies;
- the words and their n-grams;
- the service words;
- the letters and other symbols, as well as their n-grams;
- the metrorhythmic features.

The distribution of words by parts of speech and relations corresponds to the syntactic level; the frequency of punctuation marks corresponds to the punctuation level. The features describing the words and their n-grams, as well as service words, are related to the lexical level. The features describing the distribution by letters and other symbols, as well as their n-grams, largely reflect the melodica of the text and describe the text at the phonetic level. In addition, the author can consciously control such a distribution to a lesser extent, unlike, for example, the words used. The metrorhythmic features can also be attributed to the phonetic level. Thus, all levels of text description are used in the study, with the exception of stylistic, but it is important to note that the description of the text using the selected set of features is not exhaustive at any level, so further work can be aimed at building the new effective features to improve the quality of text description at various levels.

In this work, the texts of poems in modern orthography were studied. It can be assumed that the study of texts in their original pre-reform orthography would be a little more productive, but it is very difficult to collect a corpus of such texts, with the exception of A.S. Pushkin's texts. Nevertheless, such a study is of unquestionable interest and may become the subject of further research. The example of such studies of prose works in the original pre-revolutionary orthography is presented in [10]; they confirm the assumption about the productivity of this approach.

3. Related Works

The works devoted to stylometry analyze a variety of sources: prose and poetic literary works, scientific articles, diploma and term papers, essays, e-mail letters, and messages on social networks. Additionally, although a significant number of works are devoted to the analysis of literary works in prose, the analysis of poems is carried out in a small number of works [11], while only a few are devoted to the analysis of poems in Russian [12].

The construction of effective stylometry features of the text largely determines the success of the entire study of both prose and poetic texts, which is why most authors pay considerable attention to this aspect. The researchers tend to believe that the phonetic level represented by symbols and their *n*-grams is extremely productive for stylometry, the value of *n* is usually a variable parameter and largely depends on the language and features of the studied texts [8]. In addition, epy symbolic *n*-grams are easily extracted, but, nevertheless, the quality of solving problems based on this group of features alone is not enough; therefore, to improve the quality, the different authors used other features describing a wide variety of structural levels of texts.

Therefore, the authors of the work [13], which was presented within the PAN competition https://pan.webis.de (accessed on 28 January 2022) by the determination of the coincidence of authorship in pairs of imitations of original texts (fanfics), built a set of classifiers that confirm or disprove an authorship based on the analysis of a number of characteristics of the text: punctuation frequencies, last words in sentences, all categories of service words, abbreviations, verb tenses, as well as adverbs of place and time.

The authors of the work [14] took into account the syntactic structure of sentences when determining authorship; they built two self-learning subnets that accepted a sequence of words and their parts of speech, respectively. The proposed model was trained on 2662 texts of various genres in English. As a result, the consideration of the syntactic structure showed excellent results: the accuracy was 92.4%.

The paper [4] describes the approach to classifying the messages in social networks, the various combinations of *n*-gram symbols and *n*-gram words at the level of parts of speech were investigated. As a result, this approach demonstrated the accuracy of 70%.

The ensemble approach based on three independent classifiers is described in [15]. The method is based on the *n*-gram model of variable length and on the polynomial logistic regression, and is used to select a classifier with better reliability. The constructed method was tested on a set of PAN-CLEF texts in English and Portuguese, but it showed very low accuracy on the texts of songs (52%), so the method was assessed as unsuitable for lyrics.

In [16], a software product StylometryRy for determining the authorship of controversial texts in Russian is described. The texts in it were presented in the form of bag-of-words. The naive Bayesian classifier, the method *k* of the nearest neighbors and the logistic regression were used as classifiers, and the minimum text length for classification was 5500 words.

Both the Burrows' Delta and its modification, the Eder's Delta, and the author's invariant, described, for example, in [17] and defined as a characteristic of the text, calculated as the percentage of the content of service words (conjunctions, prepositions, and particles—55 words in total for the Russian language) in the text, are aimed at building the estimates of the content and the number of service words. The author of [17] notes that these characteristics are very effective for stylometry, but only within the framework of sufficiently large texts; they should be used for texts from 1000–2000 words.

The aim of the work [1] is to evaluate the effectiveness of the usage of stylometry as an alternative method of authentication of users of information systems. The authors suggested to use the descriptions of the images from users as a backup authentication system. The users were asked to describe the photos on four topics: ocean, desert, sky, and green landscapes. For the attribution of users, their writing styles are determined, which include vocabulary selection, phrasal pauses, word selection, and other features. All descriptions must be at least 200 words long. If the user forgets the account data, the backup authentication system provides another image containing the same thematic data. The new and previous descriptions are presented using a bag-of-words (BOW), after which the feature vectors are compared using the cosine similarity. If the score is above the threshold level, the author of the text will be authenticated. The model based on the support vector machine method was also built. The control experiment demonstrated an accuracy of about 73%.

The article [18] examines the approaches to determining the authorship of texts in Russian, as well as the advantages and disadvantages of these approaches. The article describes the attribution of texts, the sizes of which vary from 1000 to 100,000 symbols, using the method of support vectors (SVM) and deep neural networks with long short-term memory (LSTM), the convolutional neural networks (CNN), and transformer networks. The results show that all the considered algorithms are suitable for solving the problem of authorship identification, but SVM shows the best accuracy. The average accuracy of the SVM reaches 96%, while the accuracy increases with increasing text size and decreases with an increasing number of authors under consideration (the number of authors in the sample varied from 2 to 50). The high accuracy is due to the carefully selected parameters of the models and the feature space, which includes statistical and semantic features, among them those which are extracted as a result of the aspect analysis. Deep neural networks are inferior to SVM in accuracy and reach only 93%. The experiments show that SVM-based methods are unstable to deliberate text anonymization.

Some authors suggest to use the pre-trained language models for stylometry. The initialization of word vectors is carried out using the Word2vec model, this approach is a popular method of increasing the productivity in the absence of a large training set of texts. The work [19] describes a series of experiments with convolutional neural networks (CNN) and trained word vectors based on the Word2vec model for sentence-level classification tasks. The vectors have a dimension of 300 and were trained using the continuous bag-of-words architecture. The words missing from the set of pre-trained words were initialized randomly. The authors show that a relatively simple CNN network with a small adjustment of hyperparameters and with the word vectors based on the pre-trained Word2vec model gives excellent results on a number of tests, for example, for tonality analysis tasks. The authors of [20] used the Doc2vec model from the GENSIM3 library for cross-thematic attribution of articles by 15 authors from *The Guardian* newspaper in English. The description of the texts included *n*-grams of symbols, *n*-grams of words, and *n*-grams of tags of parts of speech; in all cases, *n* varied from 1 to 5. After the presentation of the training data in the form of *n*-grams, Doc2vec was applied to them to obtain a numerical description of the training documents. The Doc2vec model allows to represent the documents in the form of continuous and dense fixed-length feature vectors. This model generates a vector representation at the document level, it is implemented by processing each document as a special word. Thus, the authors of the work describe both the syntax and the semantics of documents, what ensures the high accuracy in attribution tasks. The logistic regression was used as a classifier. The experimental results show that the proposed model is superior to traditional Doc2vec descriptions based on unigrams of words. However, even here the authors of the work note that the greatest accuracy (at least 90%) is achieved on large texts.

A number of works attempt to explore the rhythmic component of texts. Thus, in [21], the authors propose to use phonological information about the tones and rhymes of the Chinese language, automatically extracted from unannotated texts, for the attribution of authorship. Support vector machines and algorithms based on random forests were used as classifiers. The article [22] gives the performance evaluation of the ProseRhythmDetector (PRD), the text rhythm analysis tool, for prose texts in English and Russian. The study was conducted on the basis of 50 English and 50 Russian literary texts written over the last two centuries and consisting of approximately 88,000 words each. The PRD tool was developed for the quantitative analysis of rhythm figures containing the repetition in their structure. The paper evaluates the accuracy of the PRD tool when detecting stylistic features describing the repetitions in sentences. The analysis of rhythmic figures helps to identify the idiolects of the authors and to draw conclusions about the uniqueness of their style and language, which is directly related to the problem of linguistic uniqueness and identification of the authors. The authors of the work note that, in this regard, the tool has shown encouraging results. The paper also discusses the typical errors in the operation of the tool, analyzes the controversial cases, and provides the recommendations for the usage of the tool to identify the author and his idiolect.

A number of authors successfully use the various graph representations to describe the texts and to extract from them the attributes of the author's style and attribution, as well as for annotation and semantics analysis. The graphs are also used to describe the universal dependencies. In [23,24], the adjacency graphs are used to analyze the text properties and to extract information about language styles. In the work [24], the prose literary texts in Polish and English are studied. In the adjacency graphs, the text is represented by a set of words and their word combinations; in this context, a language is understood as a complex system of higher levels that cannot be reduced to the sum of the elements involved, as, for example, the meanings of individual words do not necessarily provide the understanding of the entire sentence. Before building an adjacency network, the text undergoes the preliminary processing: annotation and unnecessary spaces are removed, all letters are converted to lowercase, and punctuation marks are replaced with special symbols so that they are then processed in the same way as words due to the fact that, from the point of view of statistical properties, the punctuation marks are no less valuable than the words, and they are the carriers of information about the language sample. In the adjacency networks, each vertex represents a single word, and the edges indicate whether words in the text occur next to each other. In the weighted version of the network, the weights represent the number of words encountered together. In this work, unlike in [23], the words are not lemmatized when constructing a network, which allows us to take into account the characteristic forms of words. A number of characteristics are calculated for each network: vertex degrees, shortest path lengths, clustering coefficients, as well as associativity and modularity coefficients. After that, the hierarchical clustering and the ensembles of decision trees are used for the studied properties. The accuracy of attribution of authorship using the described approach exceeded 90%, but it should be noted that the work investigated novels, that is, texts of considerable size.

The paper [25] describes the attribution method based on convolutional neural networks (CNNs) for symbols and their bigrams, supplemented by numerical characteristics describing the features of discourse, in order to improve the accuracy. The complexity of this approach is explained by the difficulties of constructing an adequate numerical description of the features of the discourse. For this purpose, in this work, the object grid model proposed earlier for the ordering of sentences (also called the entity grid model) was used, allowing the authors to track the grammatical relationships of the main entities. For a more productive description, the grammatical relations in this work are replaced by rhetorical structures (RS), which describe already semantic relations between fragments of sentences. The theory of rhetorical structures (RST) is intended to describe texts; it assumes a set of relations that can be observed in texts: antithesis, background, circumstance, concession, condition, and others. The table of relations is presented, for example, in https://www.sfu.ca/rst/01intro/intro.html (accessed on 28 January 2022), these relations are asymmetric and are represented by arrows of an oriented graph. The nodes of such a graph are the fragments of sentences denoting the subject (nucleus) and object (satellite) of relations. The description of the discourse structure using RST showed the excellent result and allowed the authors to increase the accuracy to 99% for 250 novels written by 50 authors.

A small number of works are devoted to the stylometry of poetic works. The work [11] describes the experiments on the analysis of poems to determine the authorship of poetic texts in Czech, German, Spanish, and English. The paper considers the Burrows' Delta, which is based on a pairwise comparison of the frequencies of frequently occurring words, as well as its modifications: Burrows' Delta, Argamon's Quadratic Delta, Smith-Aldridge's Cosine Delta, and also a classifier based on the Support Vector Machine.

The work [12] is devoted to comparing the structure of reasoning in the poetic texts of three authors: A.K. Tolstoy (two periods of creativity), K.K. Sluchevsky (two periods of creativity), and I.F. Annensky. The paper presents a logical analogue of rhetorical structures, supplemented by some relations, and demonstrates its usage for the analysis of reasoning, as well as the types of constructions and approaches preferred by the authors to the construction of discourse, and their change over time and independence from the genre were identified.

The most studies in the field of stylometry focus on the quantitative indicators for assessing the quality of models [8], while the interpretation of the results of the study in most cases remains ignored. The explanation of the decision of the classifiers would make it possible to speak with greater certainty about the features of the author's style. This work is intended to partially compensate for this disadvantage.

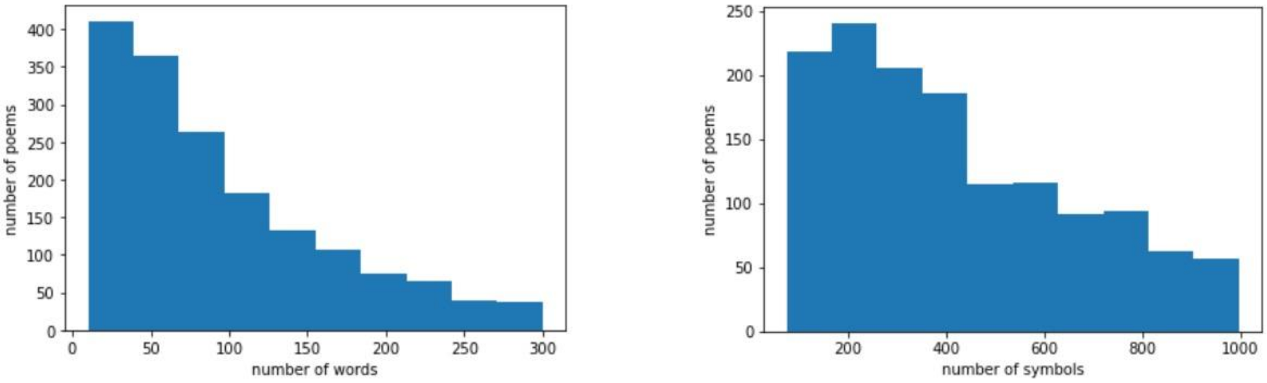
4. Data Preparation and Machine Learning Models

The task is characterized by a small amount of data, which includes a poem by A.S. Pushkin outside the lyceum period. Even taking into account this nuance, the number of poems by A.S. Pushkin in the data set was more than 34%. The poems of the other poets are presented as complete collections from the sites http://feb-web.ru/ (accessed on 28 January 2022) and https://rvb.ru/ (accessed on 28 January 2022). The description of the authors and the number of their poems in the dataset are presented in Table 1. In total, the data set consisted of 1966 poems. The works consisting of less than four lines were excluded from the set, as well as the quatrains with punctuations that are not located at the end of the poem, and with a significant number of foreign words, as this prevented the calculation of metrorhythmic characteristics.

Table 1. Data set of poems by poets of the Pushkin's era.

Author	Number of Poems	Part in the Dataset
A.S. Pushkin	685	0.3467
V.A. Zhukovsky	379	0.6533
P.A. Vyazemsky	268	
E.A. Boratynsky	197	
A.A. Delvig	194	
K.N. Batyushkov	109	
D.V. Davydov	77	
N.I. Gnedich	67	

Figure 1a shows the distribution of poems by the number of words (up to 300 words), 85% of such poems in the sample, and Figure 1b shows the same by the distribution of symbols (up to 1000); such poems in the sample are slightly more than 70%. This avenue of study will allow us to assess the effect of the size of poems on the quality of classification.



(**a**) (**b**)

Figure 1. Data set description: (**a**) The histogram of the distribution of verses by the number of words; (**b**) The histogram of the distribution of verses by the number of symbols.

Obviously, the small size of the verse limits the possibility of its accurate numerical description and makes the classification difficult, so an additional data set was built which included only poems consisting of at least 16 lines, assuming that such a data set would show better classification accuracy; the description of this set is presented in Table 2, and A. S. Pushkin's poems make up 27% of it. This set consists of only 1162 poems.

Table 2. Data set of poems by poets of the Pushkin era.

Author	Number of Poems	Part in the Dataset
A.S. Pushkin	322	0.277
V.A. Zhukovsky	261	0.723
P.A. Vyazemsky	190	
A.A. Delvig	118	
E.A. Boratynsky	106	
K.N. Batyushkov	88	
N.I. Gnedich	46	
D.V. Davydov	31	

To describe the structure of the sentences, the universal dependencies were used, which were extracted using the UDpipe2 package https://ufal.mff.cuni.cz/udpipe/2 (accessed on 28 January 2022), accessible via REST IP. Universal Dependencies (UD) is a structure for the graphical representation (agreed annotation) of grammar, including a description of parts of speech, morphological features, and syntactic dependencies, in different natural languages [26]. The dependency is a binary asymmetric relation, which is represented in diagrams by an arrow from the main to the dependent word, also to the main word of the dependent expression, if it itself is a multiword element. The sentences are represented as trees whose edges are relations, with the predicate root as the root. The dependencies are described using grammatical relationship labels. Figure 2 shows the parsing tree for the last sentence of Shakespeare's sonnet 98, obtained using the english-ewt-ud-2.6-200830 model of the UDpipe2 package. In this Figure, the relationships are signed with blue words in lowercase, and the parts of speech are green in uppercase. The root node is the main word of the phrase; it is usually a predicate and is represented by a verb, but it can also be an adjective, an adverb, etc. The display of grammatical relations between words is a distinctive feature of UD. In our work, the model russian-gsd-ud-2.6-200830 for the Russian language was used, words in initial forms, parts of speech, and relations were extracted.

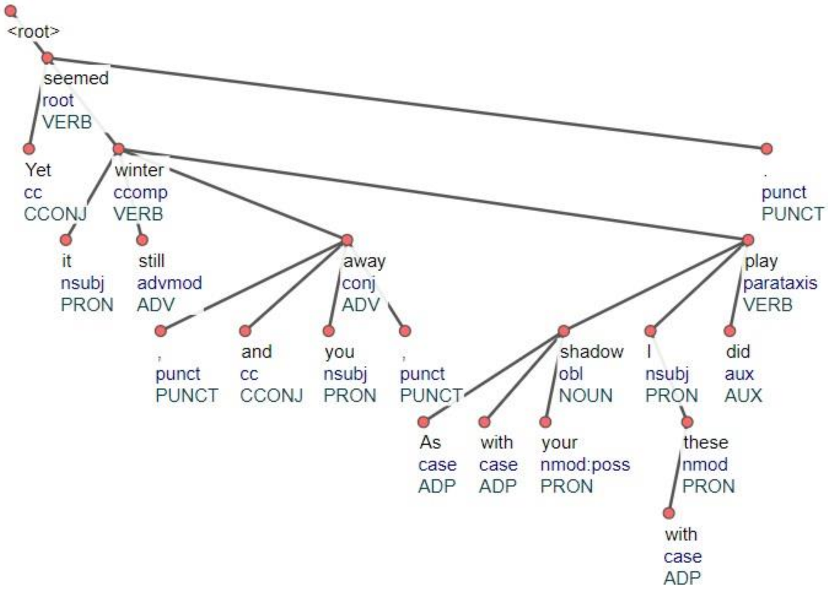


Figure 2. The syntax tree for a sentence from Shakespeare's Sonnet 98.

The TF-IDF (TF—term frequency, IDF—inverse document frequency) text model was used to obtain the "features-objects" matrix. The ensemble methods based on decision trees used for the classification include extra-trees classifier (ET), random forest classifier (RF), AdaBoost classifier (AB), gradient boosting for classification (GB), CatBoost classifier (CB), multi-layer perceptron classifier (MLP), the support vector classification (SVC), and logistic regression classifier (RF). The result tables show only the best results. The list of methods and classes implementing them are presented in Table 3. At the first stage, the classification by groups of features was performed separately for each group, then a conclusion was given about the effectiveness of the group. At the second stage, the most effective groups of features were combined, and the classification was carried out according to several groups of features. The task of the second stage was to build the most accurate classification, as well as to identify specific features that most determine the author's style.

All the given estimates are calculated on the basis of cross-validation for five blocks with the stratification by the answers. The part of the test sample is 25%. The tables show the maximum numbers of features, as well as the significances of the features which are obtained for the entire data set as a whole. It should be noted that at the stages of cross-validation, the smaller quantity of features is possible when classifying by words, service words, and also to a lesser extent by letters and symbols.

Table 3. The list of methods and classes implementing them.

Abbreviated Name of the Method	Library Name	Class Name
ET	scikit-learn	ensemble.ExtraTreesClassifier
RF	scikit-learn	ensemble.RandomForestClassifier
AB	scikit-learn	ensemble.AdaBoostClassifier
GB	scikit-learn	ensemble.GradientBoostingClassifier
CB	CatBoost	CatBoostClissifier
LR	scikit-learn	linear_model.LogisticRegression
MLP	scikit-learn	neural_network.MLPClassifier
SVC	scikit-learn	svm.SVC

5. Classification by Groups of Features

5.1. Classification Based on Parts of Speech and Relationships

To construct the classifications, the distributions by parts of speech and relations were calculated separately, after which these subgroups of features were combined. The classification by parts of speech was carried out on the basis of 16 features, and by relations on the basis of 38 features; the results are presented in Table 4. The area under curve of receiver operating characteristic (AUC ROC) for this group is only 0.7282. The best results in this group were obtained by the support vector machine. The estimation of the accuracy of classification by parts of speech is noticeably lower than the accuracy calculated by relations, and the combining subgroups of features does not allow increasing accuracy, therefore, when combining groups of features in the future, only relations will be used. It should be noted that the parts of speech and relations are naturally connected, as certain parts of speech can act as the object and subject of relations, although a certain set of options is still present here, so the features in these subgroups cannot be considered independent.

Table 4. The results of classifications by parts of speech and relations for the main data set.

Features	Method	AUC ROC	Balanced Accuracy	F1
Parts of speech	ET	0.6842	0.6365	0.5791
Relationships	SVC	0.7282	0.6627	0.5957
Parts of speech and relationships	SVC	0.7195	0.6492	0.5902

The calculations on a data set that did not include the poems of less than 16 lines gave a slight increase in the AUC ROC to 0.7355.

As, in this case, the quality of classification leaves much to be desired, the effectiveness of individual features is not discussed.

5.2. Classification Based on Punctuation Marks

For the classification based on punctuation marks, 11 features were used—the number of occurrences of each punctuation mark related to the number of words. Exclamation and question marks with two dots were considered separate from one-dot exclamation and question marks. A sequence of more than three dots was considered a separate feature (in poetic works, a row of dots usually marks a missing line). Let us note that the rules of Russian spelling have changed over time, but as the exclamation and question marks are indicators of the emotional coloring of the text, these marks can be attributed to the essential stylistic features of the text.

The results of the classification of the base of punctuation marks are presented in Table 5. Despite the small number of the features, this classification demonstrates the noticeably better accuracy indicators than the classification by parts of speech and relationships. The best results in performing this classification were demonstrated by ensemble methods based on decision trees and logistic regression. It should be noted that the most effective feature among punctuation marks was an exclamation mark. The classification of data from a set of poems with a length of at least 16 lines allowed us to obtain a better AUC ROC score of 0.7779, the best AUC ROC score for the base set was 0.7635, but the balanced accuracy is still low.

Table 5. Results of classifications by punctuation marks for the main data set.

Data Set	Method	AUC ROC	Balanced Accuracy	F1
All poems	LR	0.7635	0.7083	0.6300
From 16 lines	RF	0.7779	0.7270	0.5977

5.3. Word-Based Classification

The set of words used by the author certainly characterizes the author's style. Therefore, within the framework of this work, the classification was built according to words in the initial form and their bigrams; the results are presented in Table 6. In the classification by words, the best result was demonstrated by the neural network, the AUC ROC value was 0.8325, but the classification by words and their bigrams showed significantly worse results. Therefore, when combining groups of features, the features will be used only for separate words.

Table 6. The result of word classifications for the main data set.

n-Grams	Number of Features	AUC ROC	Balanced Accuracy	F1
1	24,136	0.8352	0.7786	0.7081
1–2	237,239	0.8126	0.7542	0.6817

By itself, the classification cannot answer the question of which words are characteristic or, conversely, not characteristic of the work of A.S. Pushkin. To construct the explanation in this paper, the Shapley method and its implementation in the form of a SHapley Additive exPlanations (SHAP) package https://shap.readthedocs.io (accessed on 28 January 2022) were used. The explanation was based on a trained CatBoost classifier. This classifier demonstrated the lower accuracy compared to neural networks (AUC ROC was 0.8072), but this accuracy turned out to be the highest among all tested ensemble methods based on decision trees. In addition, the SHAP package allows us to build an explanation for all test cases only for this group of methods.

To glean an idea of which features are most important for the constructed model, the SHAP values were constructed for each feature of each example from the source set of poems. The constructed explanation is shown in the Figure 3. The features are sorted in descending order of SHAP values, which characterizes the importance of them. The SHAP values for each test case are represented on the diagram by dots of different colors in order to show the influence of each object on the output data of the model. The dots to the left of the central vertical line represent class 0; in our case, these are other poets. Those to the right—class 1—represent the works of A.S. Pushkin. The color represents the value of the feature: red is high, blue is low. The thickness of the lines on the diagram is proportional to the number of observation points with an abscissa value.

When constructing an explanation, it is not difficult to notice that among the significant features there are many short words: pronouns and adverbs, as well as service parts of speech: prepositions, particles, and conjunctions; significantly fewer nouns and adjectives and even fewer verbs. Analyzing the work of A.S. Pushkin, it should be noted that, for his poems, the prepositions are less characteristic than for other poets, with the exception of only the preposition меж. On the other hand, the other poets whose literary work we have studied are more inclined to use the pronouns than Pushkin. The Pushkin's literary work is distinguished by a special attitude to the word свобода and, to a lesser extent, ночь, мгла, and брат; the rest of the nouns included in the 50 important features characterize the rest of the authors. Almost all significant adjectives, with the exception of one (земной), are significant for Pushkin: резвый, пустой, молодой, печальный, дальний, and others. Additionally, the poet is not characterized by the usage of the verbs быть, дать, and петь, but it is characteristic of the usage of the verb оставить, as well as adverbs наконец and вновь. In addition, Pushkin is less inclined than other authors to use the interjection ах (Figure 3).

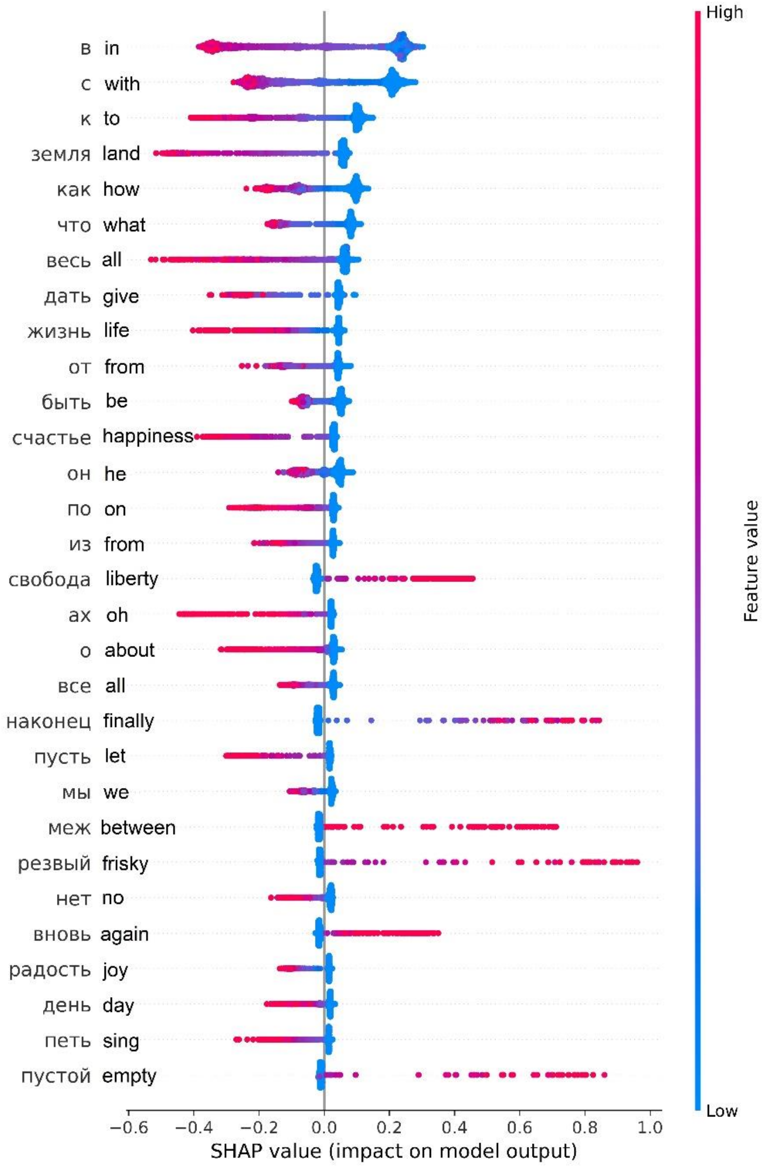


Figure 3. Distribution of the influence of the features on the results of classification by words in the initial form.

5.4. Classification Based on Service Words

In certain works [11,13,17], it was noted that the service words, namely prepositions, conjunctions, particles, and interjections, are important stylometrical features. The results of the classification of texts by words show that the pronouns should also be attributed to important indicators. Therefore, in this work, the classification was carried out on the basis of pronouns and all service words without taking into account their length. The usage of the service words, as well as sounds, is to a lesser extent the object of the author's will. Taking into account only service words allows us to ascertain information about the stylistic features of the texts of a particular author, without taking into account the semantic content of the works. With sufficient classification accuracy, this approach would be preferable to the previous version based on all words. Table 7 shows the results of classification by prepositions, conjunctions, interjections, and particles, and the results of this classification are significantly worse than when classifying by all words. The classification by words has shown that the pronouns are also significant features, but adding pronouns to the already listed parts of speech does not improve the quality of classification. The classification by one- and two-letter words shows worse results than the two previously mentioned classifications, but the classification by words of at least four letters gives the best results among all classifications in this group. Nevertheless, all the presented variants demonstrate less accuracy than the classification for all words; this can be explained by the small size of the poetic works. The paper [17] mentions that this approach is effective for the works from 1000–2000 words.

Table 7. The classification results by service and short words for the main data set.

Data	Method	AUC ROC	Balanced Accuracy	F1
prepositions, conjunctions, interjections, and particles	LR	0.7102	0.6733	0.6037
prepositions, conjunctions, interjections, particles, and pronouns	LR	0.6985	0.6038	0.5636
all words, no more than three letters long	RF	0.6701	0.6291	0.5721
all words, no more than five letters long	LR	0.7434	0.6932	0.6212

5.5. Symbol-Based Classification

In this work, the features were formed on the basis of single characters, as well as their 2 g, 3 g, and 4 g, and their various combinations. As the endings of lines have a special meaning for poems, the line feed was also taken into account. The final results for this group are presented in Table 8. The best result for single characters was obtained using the ExtraTreesClassifier method, the AUC ROC value was 0.8177. Further, the features were formed on the basis of single characters and 2 g and only 2 g; the AUC ROC value was higher when classified on the basis of only 2 g. As the classification based on bigrams is more accurate, and the addition of individual symbols to bigrams did not lead to an increase in the quality of classification, it was decided to abandon the usage of single symbols as features in the future. It should be noted that, with the increase in the number of features, the classifiers based on neural networks and the support vector method began to show significantly better results.

Next, the classifications based on 2 and 3 g and only 3 g were considered. In both variants, there is a significantly greater number of features. The comparison of these two classifiers shows that the best accuracy is demonstrated by a classifier based on 2 and 3 g using the support vector method, in this case the AUC ROC value is 0.8819, and the balanced accuracy exceeds 80%. It can be concluded that 2 g are the important features for classification; therefore, for the next attempt to improve the quality of classification, 2, 3, and 4 g were used together. As a result, the best classifier in this group was obtained, the AUC ROC value for it was already 0.8886, but at the same time the number of features increased to 64495. A further increase in the number of n-grams did not lead to an increase in accuracy.

The classification based on a dataset with the poems of at least 16 lines shows even better results, they are presented in Table 9. Here, for the best case of classification, the AUC ROC increased by more than 2% compared to the previous case and amounted to 0.9161; the best indicators are also demonstrated by balanced accuracy and F1.

The Figure 4 describing the importance of features for classification by 2, 3, and 4 g of characters using the Shepley method is shown in the figure, for clarity, the spaces on it are replaced by underlines, and the line feed is replaced by ENTER_. The analysis of n-grams of characters shows that Pushkin's poems are not characterized by the usage of an exclamation mark both at the end of the line and in the middle (that is, if the exclamation mark is not followed by a line translation), but it is very characteristic to use a dot and an ellipsis at the end of a sentence and at the end of a line, as well as a dash at the end of the line. Pushkin's poetic style is also characterized by the usage of words ending in -ый and -ные and lines in -ый. This statement is in good agreement with the result established by the classification according to words, stating that Pushkin used adjectives more than other poets. In addition, the presented diagram confirms the statement that the widespread usage of the prepositions characterizes the work of the other poets in our sample to a greater extent, especially the prepositions к, в, and с. It also follows from the presented diagram that Pushkin's poems are characterized by the usage of letter combinations ленн (primarily participles with the corresponding suffix) and words beginning with бра (as part of the roots of the words брат and брань). The letter combination ленн in the vast majority of cases is part of the participle (there are more than 100 of them): исправленный, усыпленный, влюбленный, изумленный, and притупленный; as well as several adverbs: бесмысленно, двусмысленно, медленно, мысленно, and умиленно; and only four nouns: Вселенна(я), пленник, пленница, and утопленник. The words beginning with the letter combination бра are often the words with the root бран: бранное, брань, бранить, браниться; as well as the words with the root брат: братоваться, братский, брат, братство, in addition, the characteristic words include the nouns брат and брага, as well as the adjective брадатый.

Table 8. The results of symbols classifications for the main data set.

n-Grams	Number of Features	Method	AUC ROC	Balanced Accuracy	F1
1	75	ET	0.8176	0.7590	0.6779
1–2	1782	GB	0.8579	0.7903	0.7234
		SVC	0.8357	0.7739	0.7033
2	1707	GB	0.8622	0.7870	0.7214
		SVC	0.8697	0.7984	0.7322
2–3	15,772	GB	0.8618	0.7936	0.7279
		SVC	0.8860	0.8138	0.7535
3	14,065	AB	0.8117	0.7537	0.6807
		MLP	0.8830	0.8088	0.7454
3–4	81,671	ET	0.8356	0.7669	0.6950
		MLP	0.8939	0.8239	0.7658
2–4	83,378	GB	0.8701	0.7971	0.7310
		MLP	0.8992	0.8262	0.7697
2–5	260,005	GB	0.8648	0.7975	0.7323
		MLP	0.8912	0.8215	0.7614

Table 9. The classification results by symbols for a data set from a poem no shorter than 16 lines.

n-Grams	Number of Features	Method	AUC ROC	Balanced Accuracy	F1
3–4	13,296	GB	0.8532	0.7770	0.6704
		SVC	0.913	0.8308	0.7432
2–4	62,178	GB	0.8544	0.7864	0.6749
		SVC	0.9161	0.8436	0.7580

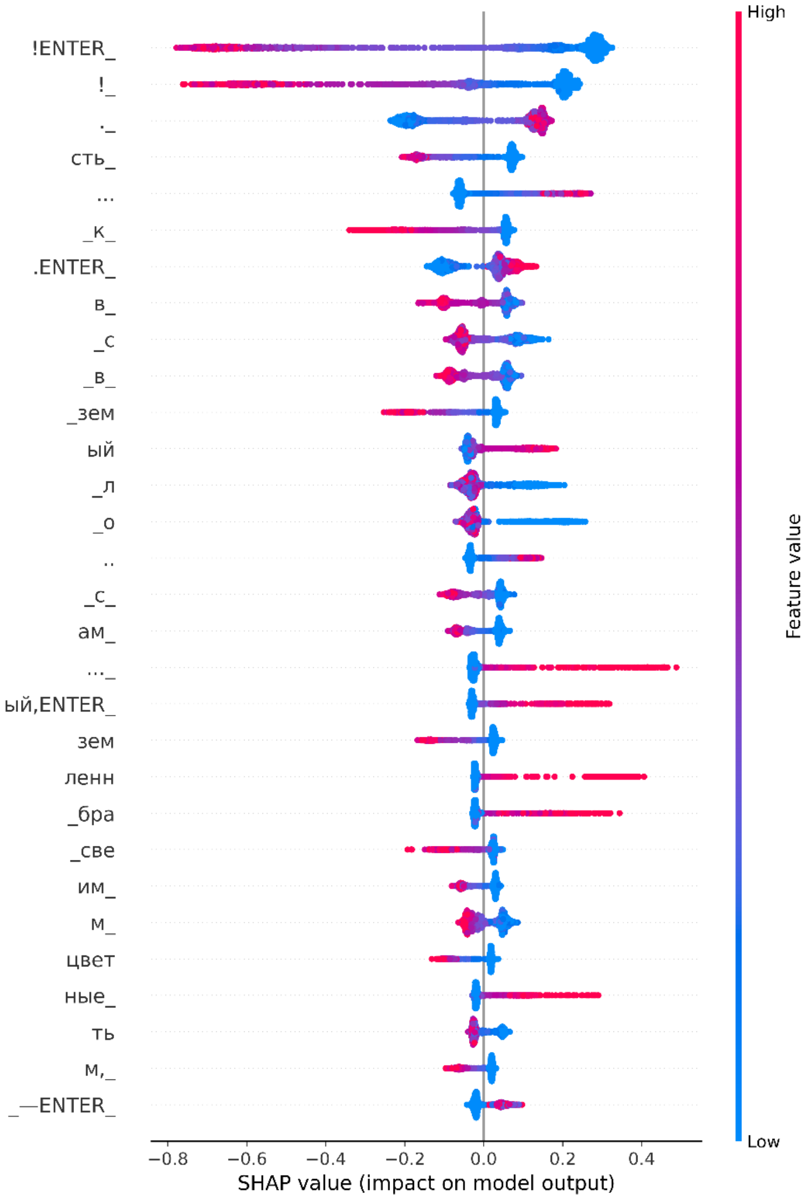


Figure 4. Distribution of the influence of the features on the results of classification by 2, 3 and 4 g.

Let us note that the classification was carried out based only on letters and spaces, but its results were noticeably worse.

5.6. Classification Based on Metrorhythmic Characteristics

The metrorhythmic features by themselves demonstrate the least accuracy, as in each subgroup, namely in meter, number of feet, and stanza, there is a variant to which most of the poems belong: 1310 poems are iambic, 918 poems belong to 4-feet meter, and 940 poems have free stanza, but in further studies when combining groups, these features turned out to be useful.

6. The Combining of the Groups of Features

After the evaluation of different groups of features, the various combinations of groups were constructed, according to the classification that was carried out. The neural networks gave the best results, so the results are given only for this model. As a result of the selection of hyperparameters, two neural networks were built with two hidden layers in each, trained according to Adam's algorithm. The results of the cross-validation are presented in Table 10.

The combination of parameters allows us to achieve a better result than for any individual group, but it should be noted that the improvement in the characteristics of the final model in comparison with the classification by symbols and their n-grams is quite small. The group of letters and their n-grams were included in the final set in two versions: 2 and 3 g, as well as 2, 3, and 4 g. Additionally, although the second option gives a little more accuracy, but it requires significantly more computing resources. In all cases, there is a greater value of completeness (about 0.9) and a lower value of accuracy. For a data set with poems with volume from 16 lines, it was not possible to obtain a significant increase in accuracy, the maximum AUC ROC on this set was 0.9240. As a result of the conducted research, the neural network was built that describes the stylistic features of A.S. Pushkin.

Table 10. The results of word classifications for the main data set.

Groups of Features	AUC ROC	Balanced Accuracy	F1
words, relations, 2–3 g of letters, metrorhythmic features, punctuation marks	0.9089	0.8478	0.7893
words, relations, 2–4 g of letters, metrorhythmic features, punctuation marks	0.9135	0.8469	0.7908

The diagram 5 describing the importance of features for classification into the groups of features using the Shapley method is shown in Figure 5, the features are supplemented with the prefixes indicating their belonging to groups: punct_—punctuation frequencies, turn_—proportions of relations, metrorhysthm_—metrorhythmic features, letter_—features by n-grams of symbols, and lemma_—features by words in the initial form. In the names of features by n-grams of characters, the spaces are still replaced by underscores, so if there are two underscores after the name of the letter_ group, then the second of them indicates a space in the n-gram. The same applies to the underscores at the end of features in the letter_ group. It is interesting to note that among the first 30 significant features, there are only two features according to words, which concerns the prepositions в and с. The main features are still the features describing the authors' preferences in punctuation marks, according to which Pushkin, unlike other authors, preferred the usage of dots and ellipses rather than exclamation marks, which distinguish the other authors. Considering the result by the words, according to which Pushkin was also not inclined to use the interjection *а*х, it can be noted that other poets of the Pushkin era were much more emotional and wrote with a more pronounced enthusiastic syllable. When the important features which describe the relations are analyzed, it should be noted that the statement that the widespread usage of functional words is not characteristic of Pushkin's work is confirmed here by the values of the features turn_case, as well as lemma_в, lemma_с, and letter__к_.

The *case* relation is a relation of case and is used for any case marking element represented by a single word; in Russian such relations connect words with prepositions. The large values of the turn_amod feature indicate the usage of modifier relations which was characteristic of A. S. Pushkin, when such relations are used the meaning of a noun or pronoun is supplemented by an adjective located both before and after the noun or pronoun to which it refers. Pushkin's commitment to the usage of adjectives was also noted when classifying by words. The features describing Pushkin's preferences in the usage of endings and letter combinations, noted when describing the classification by n-grams of characters, are also present among the important features in the final classification. The meter is also included in the set of significant features; it is represented in green on the diagram because it is categorical, not numerical data, and therefore their numerical values cannot be displayed on the diagram.

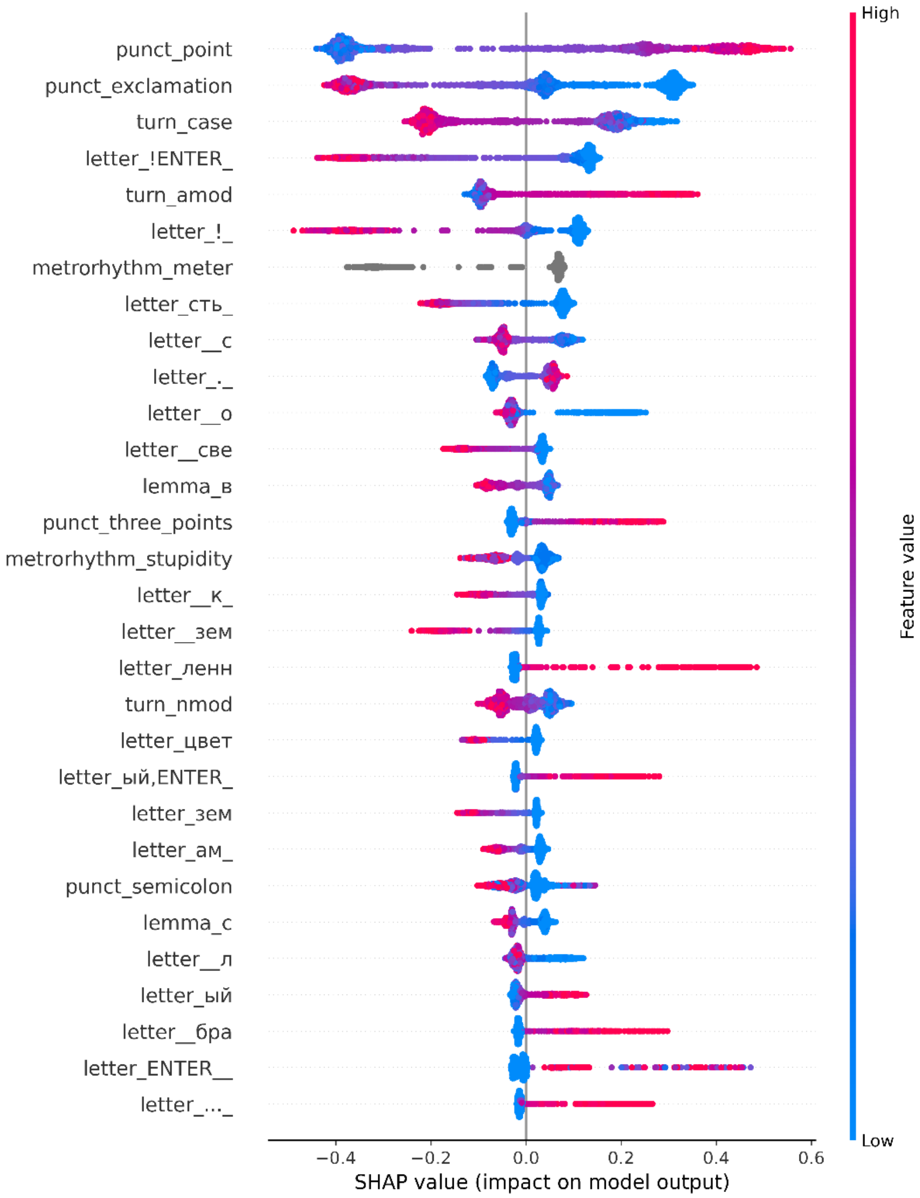


Figure 5. The distribution of the influence of features on the results of classification by feature groups.

In the main data set, the poetic works of A.S. Pushkin, written since 1823, were used. To test the constructed model, retrained on the entire data set, 50 poems by A. S. Pushkin, written in 1821 and 1822, as well as 46 poems by D. V. Venevitinov, were used. As a result, all of Pushkin's poems were classified correctly, but 13 of Venevitinov's poems were attributed to Pushkin. Thus, the F1 measure for the validation sample was 0.8849.

For the model trained on poems of at least 16 lines, all 26 of Pushkin's poems of at least 16 lines were classified correctly, and 7 of Venevitinov's 39 poems were incorrectly classified. Thus, this model shows greater accuracy, but less completeness, that is, in some cases it can attribute poems by other poets to Pushkin, but the ownership of poems by Pushkin determines almost certainly.

7. Conclusions

In this paper, the importance of groups of features of poetic texts for determining the author's style was evaluated. The most effective classification according to the considered groups of features was obtained by 2, 3, and 4 g of symbols; the classification only by letters and spaces loses to the classification by symbols by accuracy. In the next place is the classification by words. All the listed classifications showed the AUC ROC more than 0.8, but at the same time the number of classification features in them was in the thousands and tens of thousands. Next in quality was the group of punctuation marks, in which there are only 11 features. It should be noted that the punctuation marks largely reflect the emotional coloring of a poetic work. This fact is also confirmed by the quality of classification by punctuation marks, but to a greater extent by the fact that the punctuation marks (often with spaces surrounding them) have become the most important features of classification by symbols, which is described in the corresponding section.

The lowest quality was demonstrated by the classifications based on relations and metrorhythmic features. It should be noted that the distribution of relations does not fully characterize the peculiar properties of sentences; the author's preference for certain language constructions does not reflect the changes in the order of words characteristic of poetic works, what is certainly important for describing the author's style. Our attempts to build the classification features based on the analysis of sentence parsing trees have not yet brought interesting results, but this topic is extremely interesting, and the work in this direction will be continued.

The best of the constructed classifications included all the selected groups of features: 2, 3, and 4 g of symbols, words, punctuation marks, relations, and metrorhythmic features, while it is obvious that there are non-empty pairwise intersections of features in these groups.

In the future, a similar study of stylometric indicators with the usage of phonetic features and also the building of classification models by the authors on the basis of pre-trained models of vector representations of words and texts Word2Vec, Text2Vec, FastText, and neural networks based on the transformers BERT and XLM are planned.

The investigation described in this work will be included in the already existing system of the analysis of poetic texts and will correspond to the general requirements for a system formed with the requirements of philologists (potential users of the system). A separate article is devoted to this question [27].

Author Contributions: Conceptualization, V.B. and O.K.; methodology, I.G.; software, I.G.; validation, V.B., O.K. and I.G.; formal analysis, V.B. and O.K.; investigation, I.G.; resources, V.B., O.K. and I.G.; data curation, I.G.; writing—original draft preparation, I.G.; writing—review and editing, O.K.; visualization, I.G.; supervision, V.B.; project administration, V.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Russian Science Foundation, grant number 19-18-00466.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sadman, N.; Gupta, K.D.; Haque, A.; Sen, S.; Poudyal, S. Stylometry as a reliable method for fallback authentication. In Proceedings of the 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Patong Beach, Phuket, Thailand, 24–27 June 2020.
2. Kwon, H.; Lee, S. Textual Backdoor Attack for the Text Classification System Security and Communication Networks. *Artif. Intell. Cyberspace Secur.* **2021**, *2021*, 2938386.
3. Kwon, H. Dual-Targeted Textfooler Attack on Text Classification Systems. *IEEE Access* **2021**, *4*, 1. [CrossRef]
4. Mamgain, S.; Balabantaray, R.C.; Das, A.K. Author Profiling: Prediction of Gender and Language Variety from Document. In Proceedings of the 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 19–21 December 2019; pp. 473–477.
5. Gasparov, M.L. Poem. In *Literary Encyclopedic Dictionary*; Kojevnikova, V.M., Nikolaeva, P.A., Eds.; Sov. Encycl.: Moscow, Russia, 1987. (In Russian)
6. Belchikov, Y.A. Verse. In *Literary Encyclopedic Dictionary*; Kojevnikova, V.M., Nikolaeva, P.A., Eds.; Sov. Encycl.: Moscow, Russia, 1987. (In Russian)
7. Anwar, D.; Bajwa, I.; Ramzan, S. Design and Implementation of a Machine Learning-Based Authorship Identification Model. *Sci. Program.* **2019**, *14*, 9431073. [CrossRef]

8. Lagutina, K.; Lagutina, N.; Boychuk, E.; Vorontsova, I.; Shliakhtina, E.; Belyaeva, O.; Paramonov, I.; Demidov, P.G. A Survey on Stylometric Text Features. In Proceedings of the 25th Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 5–8 November 2019; pp. 184–195.
9. Batura, T.W. Formal methods of attribution of texts and their implementation in software products. *Softw. Prod. Syst.* **2013**, *4*, 286–295. (In Russian)
10. Kozhemyakina, O.Y.; Tagirova, E.P. The translation algorithm from pre-reform spelling into modern spelling, taking into account the morphology of words. *J. Phys. Conf. Ser.* **2019**, *1405*, 012010. [CrossRef]
11. Plecháč, P.; Bobenhausen, K.; Hammerich, B. Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry. *Studia Metr. Poet.* **2018**, *5*, 29–54. [CrossRef]
12. Timofeeva, M. Comparative Analysis of Reasoning in Russian Classic Poetry. *Appl. Sci.* **2021**, *11*, 8665. [CrossRef]
13. Halvani, O.; Graner, L.; Regev, R. Cross-domain authorship verification based on topic agnostic features. In *Proceedings of the Working Notes of CLEF, Thessaloniki, Greece, 22–25 September 2020*; Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A.F., Eds.; 2020. Available online: http://ceur-ws.org/Vol-2696/paper_114.pdf (accessed on 9 December 2021).
14. Jafariakinabad, F.; Hua, K.A. A Self–Supervised Representation Learning of Sentence Structure for Authorship Attribution. *arXiv* **2020**. Available online: https://arxiv.org/abs/2010.06786 (accessed on 9 December 2021). [CrossRef]
15. Custodio, J.E.; Paraboni, I. An ensemble approach to cross-domain authorship attribution. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Lugano, Switzerland, 9–12 September 2019; pp. 201–212.
16. Chashchin, S.V. Application of "supervised" machine learning methods for text attribution: Individual approaches and intermediate results in identifying authors of Russian-language texts. *Probl. Criminol. Forensic Sci. Forensic Exam* **2018**, *1*, 139–147.
17. Batura, T.W. Formal methods of attribution of texts. *Vestn. NGU. Ser. Ser. Inf. Technol. Inf.* **2012**, *10*, 81–94. (In Russian). Available online: https://lib.nsu.ru/xmlui/handle/nsu/258 (accessed on 8 December 2021).
18. Romanov, A.; Kurtukova, A.; Shelupanov, A.; Fedotova, A.; Goncharov, V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet* **2021**, *13*, 3. [CrossRef]
19. Yoon, K. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**. Available online: https://arxiv.org/pdf/1408.5882.pdf (accessed on 8 December 2021).
20. Barlas, G.; Stamatatos, E. Cross-Domain Authorship Attribution Using Pre-Trained Language Models. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Neos Marmaras, Greece, 5–7 June 2020; pp. 255–266. Available online: https://link.springer.com/content/pdf/10.1007%2F978-3-030-49161-1_22.pdf (accessed on 8 December 2021).
21. Hou, R.; Huang, C.R. Robust stylometric analysis and author attribution based on tones and rimes. *Nat. Lang. Eng.* **2020**, *26*, 49–71. [CrossRef]
22. Boychuk, E.; Lagutina, K.; Vorontsova, I.; Mishenkina, E.; Belyayeva, O. Evaluating the Performance of a New Text Rhythm Analysis Tool. *Engl. Stud. NBU* **2020**, *6*, 217–232. [CrossRef]
23. Amancio, D.R. A complex network approach to stylometry. *PLoS ONE* **2015**, *10*, e0136076. [CrossRef] [PubMed]
24. Stanisz, T.; Kwapien, J.; Drozdz, S. Linguistic data mining with complex networks: A stylometric-oriented approach. *Inf. Sci.* **2019**, *482*, 301–320. [CrossRef]
25. Ferracane, E.; Wang, S.; Mooney, R. Leveraging discourse information effectively for authorship attribution. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 27 November–1 December 2017; Volume 1, pp. 584–593.
26. Marneffe, M.-C.; Manning, C.; Nivre, J.; Zeman, D. Universal Dependencies. *Comput. Linguist.* **2021**, *47*, 255–308. [CrossRef]
27. Barakhnin, V.B.; Kozhemyakina, O.Y.; Mukhamediev, R.I.; Borzilova, Y.S.; Yakunin, K.O. The design of the structure of the software system for processing text document corpus. *Bus. Inform.* **2019**, *13*, 60–72. [CrossRef]